Variable Temptations and Black Mark Reputations

HARVARD Kennedy School
JOHN F. KENNEDY SCHOOL OF GOVERNMENT

Variable Temptations and Black Mark Reputations

Faculty Research Working Paper Series

Christina Aperjis

HP Labs

Yali Miao

Jane Street Capital

Richard J. Zeckhauser

Harvard Kennedy School

March 2011
RWP11-020

Variable Temptations and Black Mark Reputations[☆]

Christina Aperjis

HP Labs, 1501 Page Mill Rd, Palo Alto, CA 94304

Yali Miao

Jane Street Capital, Roppongi 6-12-4, Minato-ku, Tokyo, Japan 106-0032

Richard J. Zeckhauser

Harvard University, 79 John F. Kennedy Street, Cambridge, MA 02138

Abstract

In a world of imperfect information, reputations often guide the sequential decisions to trust and to reward trust. We consider two-player situations, where one player — the truster — decides whether to trust, and the other player — the temptee — has a temptation to betray when trusted. The strength of the temptation to betray varies from encounter to encounter. We refer to a recorded betrayal as a black mark and focus on mechanisms that only reveal the number of black marks of a temptee. We show that the greater the number of black marks, the less likely the temptee is to betray. We then study the different equilibria that emerge, depending on which side of the market has the ability to specify the equilibrium. In closing, we generalize to cases where the number of encounters is also recorded.

Keywords: Game Theory, Trust, Reputation

1. Introduction

In a typical business transaction, one or both parties have the potential to betray. A supplier can produce low-quality goods; a debtor can default; an employee can steal; or a contractor can break the deal. Betrayals are often avoided because temptations are modest or nonexistent. But even when temptations are significant, reputations can keep untrustworthy behavior in line. Thus, betrayal is deterred, lest we lose future business with others, find ourselves without future credit or facing higher interest rates from any lender, or have great difficulty

[☆] We are grateful to Vincent Crawford, John H. Lindsey II, Ramesh Johari, Paul Resnick, Ashin D. Shah and Peter Zhang for helpful comments. This work was partially supported by a grant from the Alfred P. Sloan Foundation, and the NSF under award IIS-0812042.

finding a job. Many economic models focus on repeat play, but often the concern with reputation comes from the broader world. Personal interactions, as between friends, present the same situation, with temptations, betrayals, and reputations all playing important roles.

Reputations are hardly sufficient statistics. They rarely tell us everything or almost everything about an individual's prior credit dealings or employment history. A typical employee reference in these litigious days is likely to be: "Joe worked here for 12 years, and there are no recorded blemishes on his record." Information on credit scores is equivalently crude. Repaying a loan counts the same whether the terms were easy or harsh. If a minimum grade point average is necessary to keep one's scholarship, it is irrelevant if one's courses are easy or hard. The Better Business Bureau[1] only has information on the number of complaints for a business, but does not provide details on each complaint. On the well-known and highly successful eBay reputation system, the summary score tells us how many positive, neutral, and negative feedbacks a seller has received, but not the highly relevant information of the prices of all the items that received negative feedback.[2]

In this paper, we study the *black mark reputation mechanism*, a mechanism where an individual's reputation is simply a tally of the number of bad ratings or complaints that she has received. In a broad range of settings, the reputation mechanism only keeps track of the number of infractions. For example, the Better Business Bureau has information on the number of complaints for a particular business, but not the number of transactions or volume of business that might have led to complaints. On the other hand, in some instances, an infraction carries weight in and of itself, and people do not think (or recognize) that the number of trials matters. This is in the spirit of criminal justice systems, where the judge learns the number of convictions in a defendant's past before sentencing, or some systems of sexual morality which look at the number of partners someone has had.

Black mark reputations approximate settings where buyers focus on the number of negatives — even if more reputation information is provided. A strongly negative rating is a black mark. The Internet is now bristling with websites where users rate firms. They include Tripadvisor (for hotels and restaurants), Angie's List (for service providers), and Yelp (for restaurants, attractions, etc.). Participants give individual feedback scores after an encounter. Most rated entities, even those with good average reputations, have some very low scores, usually from some disastrous encounters. Some potential buyers focus on the

[1] www.bbb.org

[2] The concern, of course, is that the seller would be generally trustworthy, but dishonest on rare occasions when describing a very high-priced item. Information regarding sold items remains available for 90 days on eBay. It is thus possible to scroll through to see the sale prices of recently sold items. This just complicates the strategy of the dishonest seller, who must take a break after doing an untrustworthy transaction at a high price. In any case, information that is prominently shown to buyers on eBay has a larger effect than information that is available but harder to find (Cabral and Hortacsu, 2010).

extreme negatives: "The restaurant lost our reservation and could not seat us;" "The plumber showed up 6 hours late, and left the place a mess." Businesses presumably know that further highly negative encounters could be extremely damaging, and will strive to improve their reservation system or their promptness.

We focus on two-player situations, where one player — the *truster* — decides whether to trust, and the other player — the *temptee* — has the temptation to betray when trusted. Behaviors differ in our model — as in real life — because *the strength of the temptation to betray varies from encounter to encounter.* The tempted players could be suppliers who might breach a contract that turns out to be too costly, contractors who might do a shoddy job if it saves a lot of effort, employees who might miss work often when other responsibilities are pressing, or spouses who might stray from marital vows given highly attractive opportunities. The strength of the temptation to betray is then assumed to vary according to some probability distribution.

Given our focus on trust, we employ the term reputation, in essence what players know about another when they decide whether to trust. However, our analysis extends to a range of punishment situations, where the more general term record would be more appropriate. When a professor turns in a plagiarizer or a policeman writes up a ticket for a speeder, neither knows the number of infractions the individual has to date. But since this information is recorded, and punishment depends on the number of infractions (e.g., two plagiarism convictions and you are expelled), recipients of black marks will behave in the manner described here. A second generalization should be mentioned at the outset. In section 8, we consider reputations that keep track of the number of trials.

This paper addresses two major questions.

(1) Are people with worse reputations more likely to betray?

(2) Do different equilibria for the treatment of reputations emerge depending on which side of the market has the ability to specify the equilibrium?

The answer to even the first question is hardly clear. Suppose we have a "two strikes and you're out" system. Will a player with one strike be more likely to betray than a player with no strikes? We show that in any pure equilibrium of a game where a temptee's reputation depends solely on the number of black marks, *the greater the number of black marks, the less likely the temptee is to betray.* Moreover, under certain probability distributions of the temptation's strength, the likelihood of betraying decreases faster when the temptee's reputation consists of a larger number of black marks.

It may seem counterintuitive that those with worse reputations would behave better. We can get some intuition by observing that since the truster's decision to trust only depends on the number of black marks of the temptee, there must be a cutoff after which he stops trusting her. The temptee's best response to such a cutoff strategy is to be more careful when she gets closer to the cutoff, or in other words, to betray less when her reputation becomes worse.

The qualitative equivalent of the situation we address is seen in the Goldman Sachs (GS) situation after its revenue was tarnished by allegedly proffering portfolios designed by famed short seller John Paulson, who wished to sell them short. GS allegedly did not reveal this highly relevant information and is currently subject to both regulatory impositions and significant lawsuits. Quite apart from any legal action, its reputation has suffered gravely. Holding GS's type fixed, it now seems much less likely that it would allow itself to engage in another ploy of this sort. Another round of such allegations could sound the death knell for GS as a leading investment house.

We address the second question by considering which pure equilibria are best for each of the two players, that is, the truster and the temptee. We demonstrate that the preferences of the two players may in general be dramatically different. However, there are cases where both the truster and the temptee prefer the same pure equilibrium, making a socially optimum available.

In order to simplify the exposition, for most of the paper we set aside any information about types and consider a pure moral hazard setting, where all players on one side of the market are identical in terms of the distribution of the temptation's strength and the payoff structure. In the pure moral hazard setting, the role of the reputation system is to deter temptees from bad behavior. However, our results still hold in settings with both moral hazard and adverse selection. Then, in addition to deterring bad behavior, the black mark reputation system may also help to sort out those who tend to get strong temptations to betray.

In closing, we consider mechanisms that track both the number of black marks and the number of encounters of a temptee, and show that the equilibrium behavior in the long run is identical to equilibrium behavior under a black mark reputation mechanism.

The remainder of the paper is organized as follows. Section 2 discusses related literature. The problem formulation is given in section 3, and a general characterization of pure equilibria is presented in section 4. Section 5 shows that in any pure equilibrium the temptee is less likely to betray when she has more black marks. Section 6 studies how the equilibrium is determined by the one who specifies it. Section 7 considers adverse selection. Section 8 considers a setting where the truster knows both the number of black marks and the number of transactions of the temptee. Section 9 concludes. All proofs are provided in Appendix A. For completeness, we discuss mixed equilibria in Appendix B.

2. Related Literature

Reputation is often studied in settings with both adverse selection and moral hazard. Reputational concerns usually incentivize the agent to act in ways that benefit her counterparties.[3] The agent is assumed to have a type that her counterparts try to infer from her past behavior (for a survey of such models see

[3] In certain situations reputational concerns may have the opposite effect (Ely et al., 2008).

Mailath and Samuelson, 2006). In this setting, an agent's reputation represents the belief that other players have about her type, and the analyses focus on sorting among types. An agent with a worse reputation is thought less likely to be of a good type and is considered more likely to betray. For instance, Sobel (1985) considers an adverse selection model, where one player (the sender) is either a friend or an enemy and the other player (the receiver) has a prior belief on the sender's type; it turns out that an enemy is less likely to lie when she has a better reputation.

By contrast, in pure moral hazard models, the agent does not have a hidden type and reputation is only used to incentivize good behavior. This approach has been taken by Dellarocas (2005) in the context of an electronic marketplace, where the seller might betray the buyer; his paper studies mixed equilibria in which the seller cheats (betrays) the buyers with some probability that decreases with her reputation. Thus, a seller with a better reputation is less likely to betray. In this paper, we do most of the analysis for a pure moral hazard model, but our results also apply to settings where both moral hazard and adverse selection come into play. Moreover, we show that agents are more likely to betray when they have better reputations in that setting.

The literature on optimal penal codes (Abreu, 1988; Lambson, 1987; Abreu et al., 1990; Athey and Bagwell, 2001) is also related to our work. An application that is often considered in this literature is maximal collusion in a repeated oligopolistic game; to support the collusive outcome, a firm that deviates is punished. A key insight of this literature is that there exists a simple punishment (that is independent of the history of the game in some sense) that is optimal. In this paper, we do not study the issue of optimal punishment. Instead, we consider settings where the truster only knows the number of black marks of the temptee and identify which equilibria are optimal for each side of the market.

Although we do not consider incentives for trusters to leave honest feedback in this paper, another extensive line of research considers how truthful feedback can be elicited. In online markets, players might undertake fake transactions in order to enhance their reputations. This stratagem is unattractive, however, if a specific relation between the reputation premium and the transaction cost holds (Bhattacharjee and Goel, 2005). On the other hand, even if fake transactions cannot be undertaken, buyers might not leave honest feedback after a transaction. Nevertheless, it is still possible to devise a scoring system that induces honest reporting of feedback (Miller et al., 2005). In this paper, we posit that trusters leave honest feedback, since they have no reason not to do so. We focus on the temptee's decision and its influence on the decision to trust. We do, however, allow for imperfect monitoring in our model, as various studies have shown that monitoring is often imperfect in practice (Bolton et al., 2009; Dellarocas and Wood, 2008; Chwelos and Dhar, 2008).

3. Problem formulation

Players in the temptation game are divided into two roles, Cs and Ds, trusters and temptees. For expository ease, those who must decide whether to trust — trusters — are males, and those who are subject to temptation — temptees — are females in our analysis.

Our framework and results apply to the following settings:

- *Short-term interactions between trusters and temptees.* In each period, there are equal numbers of temptees and trusters, and each truster is randomly matched up with a temptee. In this case, one contracts with another party for just one period, and then moves on.

- *Long-term interactions between one truster and a large number of temptees.* In this setting, the truster interacts with multiple temptees at the same time (in each period).[4] For instance, the truster can be a big employer interacting with multiple employees, a university interacting with many students, or a state interacting with a large number of citizens.

Throughout the paper we consider both situations simultaneously. We thus occasionally say "the truster(s)" to refer to the set of trusters in the setting of short-term interactions and to the truster in the setting of long-term interactions between one truster and a large number of temptees.

In every period, C decides whether to choose "trust" or "safe." If C plays trust, then D can play "reward" or "betray." If D rewards, then D and C both get a unit payoff. If D chooses to betray, then D will get a $(1+x)$ of payoff, where x is the strength of the temptation to betray. Its magnitude is the realization of a random variable X. In each round, prior to choosing whether she will reward or betray, D learns her x for that round, namely her level of temptation. The value of x is and remains unknown to C. Thus, no information on the strength of the incentive to betray is ever part of a D's reputation. If D chooses to betray, then C will get a payoff of -1. Figure 1 shows the extensive form representation of a one-period interaction between C and D, where C's choices are circles and D's are squares, and C's payoff is listed first. The analysis remains qualitatively the same if C gets a payoff of $-y$ when D betrays, rather than -1, though of course the parameter values at equilibria will shift. We note that the scaling of these payoffs is arbitrary. There is no implied interpersonal comparison. For example, in dollar value C may gain far more than D when each goes from 0 to 1.

We refer to a recorded betrayal of D as a black mark. The number of black marks a D has received is known to a C when he encounters her. We allow for imperfect recording; that is, recording rewards as betrayals and betrayals as rewards. In particular, if D betrays, the number of black marks increases by 1 with probability $1 - r$ and remains the same with probability r. If D rewards, then the number of black marks remains the same with probability $1 - q$ and increases by 1 with probability q. Perfect monitoring is a special case with $r = q = 0$. We refer to the number of recorded betrayals (or black marks), b, as

[4] Our results can be extended (but do not directly apply) for settings where each truster has a long-term relationship with one temptee.

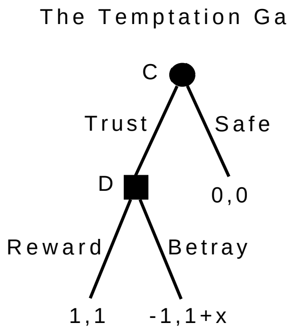

Figure 1: Extensive form representation of one-period interaction between C and D. (C's payoff listed first.)

the *reputation* of D. When monitoring is perfect, b equals the actual number of betrayals. In general, however, it may differ.

At each round, each D has a certain probability of surviving to the next period, s. We leave aside discounting, except as it arises through D's survival concerns. Absent discounting, the survival rate for Cs turns out to be nonmaterial. After each round, if a D dies, she will be replaced by another D who enters with a blank reputation record. If the reputation of a temptee ensures she will no longer be trusted (in our language she gets expelled), then she is not trusted until she dies (and is replaced by a new temptee with blank reputation only after she dies). We further assume that all players are risk-neutral. The goal of a D is to maximize her expected payoff until she dies or is expelled. The goal of a C is to maximize his expected payoff each period, because we take a steady-state approach.

Key notation introduced in this and subsequent sections is summarized in Table 1. As we discuss in Section 4, we make some general assumptions to rule out settings with uninteresting equilibria. In particular, we are assuming that the random variable X has a finite mean and a strictly positive median, that the imperfect monitoring probabilities are not too large (in particular, $r + q < 1$), and that the survival probability of the temptee is strictly between 0 and 1.

4. Characterization of Pure Equilibria

In this section we characterize pure equilibrium. In the following sections we use this characterization to show properties of these equilibria. Mixed equilibria are analyzed in Appendix B.

For the game to be in equilibrium, at each round, both C and D must have no incentive to deviate from the strategies defining the equilibrium. C's strategy consists of whether he trusts D as a function of D's reputation. D's strategy determines whether she rewards as a function of her reputation b and her realization of X in that period.

Notation	Definition
X	strength of temptation to betray
r, q	imperfect monitoring probabilities
s	survival probability of the temptee (player D)
b	number of (imperfectly) recorded betrayals, or black marks
b^*	number of black marks at which C stops trusting D
$v(b)$	D's maximum expected infinite horizon payoff at b black marks
$x^*(b)$	threshold on X below which D rewards, and above which she betrays
$w(b)$	probability that D betrays when she has b black marks

Table 1: Notation used in the paper.

For a fixed strategy of C, let b^* be the minimum number of betrayals at which C does not trust D. That is, C trusts when $b < b^*$ and does not trust when $b = b^*$. Since C does not trust D at b^*, D will never have more than b^* black marks. We thus refer to b^* as the *cutoff* at which C stops trusting D.

We first consider the best response of player D when C uses a cutoff b^*. Let $v(b)$ be the maximum expected infinite horizon payoff to D when her reputation consists of b black marks. We note that if monitoring is imperfect (i.e., $r+q > 0$), then b may be different than the actual number of D's betrayals. Since the cutoff is b^*, C will never trust D once her reputation becomes b^*, and thus

$$v(b^*) = 0. \tag{1}$$

For $b \in \{1, ..., b^* - 1\}$, $v(b)$ is equal to

$$\mathbb{E}[\max\{1 + X + s((1-r) \cdot v(b+1) + r \cdot v(b)), 1 + s((1-q) \cdot v(b) + q \cdot v(b+1))\}].$$

In particular, given that the realization of the random variable X is x, D chooses the action that maximizes her expected payoff. Should she choose to betray, her expected payoff is $1+x+s((1-r)\cdot v(b+1)+r\cdot v(b))$, since she receives $1+x$ now and her reputation deteriorates to $b+1$ black marks with probability $1-r$ and remains the same (i.e., equal to b black marks) with probability r. On the other hand, if D chooses to reward, her expected payoff is $1+s((1-q)\cdot v(b)+q\cdot v(b+1))$, since she receives 1 now and her reputation remains the same (equal to b) with probability $1-q$ and deteriorates to $b+1$ black marks with probability q. The temptee selects to reward or betray depending on which action gives her the largest expected payoff.

Straightforward calculations show that $v(b)$ is described by the following dynamic program:

$$v(b) = 1+s(1-q)\cdot v(b)+sq\cdot v(b+1)+\mathbb{E}[(X-s(1-r-q)\cdot(v(b)-v(b+1)))^+], \tag{2}$$

where $y^+ \equiv \max(y, 0)$ is the positive part of y. Let

$$x^*(b) = s(1 - r - q) \cdot (v(b) - v(b + 1)). \tag{3}$$

Equation (2) implies that it is optimal for D to reward if $X < x^*(b)$ and betray if $X > x^*(b)$. D is indifferent between rewarding and betraying when $X = x^*(b)$. Thus, both rewarding when $X < x^*(b)$ and rewarding when $X \leq x^*(b)$ are optimal for D. For simplicity, we will assume that D chooses to reward if and only if $X < x^*(b)$ in this case.[5] This simplifies the presentation because now the set $\{x^*(b), b = 0, 1, ..., b^*\}$ characterizes the best response of D. However, this assumption is not essential for our results.

Substituting (3) into (2) we conclude that

$$(1 - s(1 - q))v(b) = sq \cdot v(b + 1) + 1 + \mathbb{E}[(X - x^*(b))^+]. \tag{4}$$

(We assume that X has a finite mean, so that $\mathbb{E}[(X - x^*(b))^+]$ is well defined.)

Since D gets strictly positive immediate payment whenever she is trusted, the value $v(b)$ is strictly decreasing for $b \leq b^*$. This is shown formally in the following lemma.

Lemma 1. *For a given cutoff $b^* < \infty$, suppose that $v(b)$ and $x^*(b)$ satisfy* (1)*,* (3)*, and* (4)*. Then $v(b)$ is strictly decreasing in $\{0, 1, ..., b^*\}$.*

If the level of imperfect monitoring is sufficiently high, namely if $r + q \geq 1$, then it is impossible to incentivize D to reward when the strength of the temptation to betray is positive (since then Lemma 1 and Equation (3) imply that $x^*(b) \leq 0$). We are not interested in studying such cases and assume that $r + q < 1$ throughout the paper. Lemma 1 then implies that $x^*(b)$ is strictly positive for $b \in \{0, 1, ..., b^* - 1\}$.

We next consider player C. Given $x^*(b)$, the expected payoff of C is

$$\mathbb{P}[X < x^*(b)] - \mathbb{P}[X \geq x^*(b)] = 2 \cdot \mathbb{P}[X < x^*(b)] - 1$$

if he trusts D; and 0 otherwise. We conclude that C trusts D if $\mathbb{P}[X < x^*(b)] > 1/2$; C does not trust D if $\mathbb{P}[X < x^*(b)] < 1/2$; and C is indifferent between trusting and not trusting if $\mathbb{P}[X < x^*(b)] = 1/2$.[6]

Let m be the median of X, i.e., m is such that $\mathbb{P}[X \leq m] \geq 1/2$ and $\mathbb{P}[X \geq m] \geq 1/2$. If $m \leq 0$, then C always trusts D at the equilibrium. We are interested in settings where this is not the case and thus assume that $m > 0$. We first observe that, if $m > 0$, then there cannot exist an equilibrium where C always trusts D. In particular, if C always trusts D, then D's best response is to

[5] In most cases, $\mathbb{P}[X = x^*(b)] = 0$ and thus it is not essential to specify what D does when $X = x^*(b)$. In particular, for a continuous distribution, $X = x^*(b)$ with probability zero. On the other hand, when the distribution is discrete, then $x^*(b)$ is usually at a point of zero mass.

[6] The number 1/2 arises because we are assuming that C's payoff is equal to -1 when D betrays. More generally, if C got a payoff of $-y$ (instead of -1) when D betrayed, our subsequent analysis would go through by replacing 1/2 by $y/(y + 1)$.

always betray whenever there is a positive temptation to betray, i.e., $x^*(b) = 0$ for all b; however, C's best response to $x^*(b) = 0$ for all b is to never trust, since $m > 0$. We conclude that $b^* < \infty$.

At an equilibrium, both C and D play a best response to the other player's strategy. In particular, a cutoff $b^* \geq 0$ and the set $\{x^*(b), b = 0, 1, ..., b^*\}$ constitute a pure equilibrium of the temptation game if the following conditions are satisfied:

1. $x^*(b)$ is a best response of D to C's strategy, i.e., there exists a function $v(b)$ such that $x^*(b)$ and $v(b)$ satisfy (1), (3), and (4)
2. b^* is a best response of C to D's strategy, i.e.,
   - $\mathbb{P}[X < x^*(b)] \geq 1/2$ for $b < b^*$
   - $\mathbb{P}[X < x^*(b^*)] \leq 1/2$

A pure equilibrium can be computed by recursively solving Equations (3) and (4) to obtain $x^*(b^*-i)$ and $v(b^*-i)$ starting from the initial condition given by (1). Then, the cutoff b^* and the computed set $\{x^*(b), b = 0, 1, ..., b^* - 1\}$ constitute an equilibrium if $\mathbb{P}[X < x^*(b)] \geq 1/2$ for $b < b^*$. On the other hand, if $\mathbb{P}[X < x^*(b)] < 1/2$ for some $b < b^*$, then there does not exist a pure equilibrium with cutoff greater or equal to b^*.

We observe that there always exists a degenerate equilibrium where C never trusts D and D never rewards, that is, $b^* = 0$. The temptation game has a non-degenerate equilibrium (i.e., with $b^* \geq 1$) if the solution y of the following equation

$$\frac{1 - s(1-q)}{s(1-r-q)} y = 1 + \mathbb{E}[(X-y)^+]$$

is greater than m. This follows by considering Equations (1), (3), and (4) for $b = b^* - 1$.

Even though $b^* < \infty$ and, therefore, C does not trust D after a finite number of black marks, there may exist equilibria at which cooperative behavior is sustained for the duration of D's lifetime. Alternatively, if we interpret D's survival probability as a discount factor and the monitoring is perfect, then cooperation can be sustained forever, along the lines of a folk theorem (Myerson, 1997). In particular, suppose that $b^* = 1$ and D always rewards when she has no black marks (i.e., $\mathbb{P}[X < x^*(b^* - 1)] = 1$). Straightforward calculations show that this is the case if the maximum possible value of X is less than $s(1-r-q)/(1-s(1-q))$. On the other hand, there may also exist equilibria with $b^* > 1$, where D may betray when she has strictly less than $b^* - 1$ black marks, but always rewards when she has $b^* - 1$ black marks. For instance, if $b^* = 2$ represents an equilibrium and the maximum possible value of X is less than $s(1-r-q)/(1-s(1-q))$, then D will betray with positive probability when she has no black marks, but will never betray when she has one black mark. Thus, in some sense, cooperation is sustained, with one betrayal. The outcome is similar for other equilibria where the permissible number of black marks is greater.

In general, if there exists an equilibrium with cutoff $b^* = k$, there also exists a pure equilibrium with cutoff $b^* = k'$, where $k' < k$ (assuming that both k and k' are positive integers). Let B^* be the maximum cutoff b^* for which there exists a (pure) equilibrium. The value of B^* depends on the random variable X, the survival probability s, and the imperfect monitoring probabilities r and q.

5. Betrayal as a Function of Reputation

In this section we consider how reputations work. We find that temptees are less likely to betray when they have bad reputations, and that (for a plausible class of distribution functions) the likelihood of betraying decreases faster when the temptee's reputation consists of a larger number of black marks. The following proposition states this result formally.

Proposition 1. *For every pure equilibrium* $(b^*, \{x^*(b), b = 1, 2, ..., b^*\})$ *of the temptation game,* $x^*(b)$ *is strictly increasing and convex in* b *for* $b \in \{0, ..., b^* - 1\}$*.*

Proposition 1 shows that the threshold $x^*(b)$ is increasing and convex in the number of black marks. What are the implications of this result on how likely player D is to betray? Let

$$w(b) \equiv \mathbb{P}[X \geq x^*(b)].$$

That is, $w(b)$ is the probability that D betrays when she has b betrayals. Moreover, let F be the cumulative distribution function of the random variable X.

The following corollary of Proposition 1 characterizes $w(b)$.

Corollary 1. *For every pure equilibrium* $(b^*, \{x^*(b), b = 1, 2, ..., b^*\})$ *of the temptation game:*

(i) $w(b)$ *is decreasing in* b *for* $b \in \{0, ..., b^* - 1\}$

(ii) if F *is linear or convex, then* $w(b)$ *is concave in* b *for* $b \in \{0, ..., b^* - 1\}$

In words, in any game where a temptee's reputation depends solely on the number of black marks, the more black marks to date, the less likely the temptee is to betray. This follows from the fact that $x^*(b)$ is increasing in b. It may seem counterintuitive that those with worse reputations would behave better. However, since the truster is using a cutoff strategy, the temptee's best response is to be more "careful" when she gets closer to the cutoff, or in other words, to betray with a smaller probability (that is, only for very large temptations) when her reputation becomes worse.

In addition to showing that there is a decrease in the likelihood of betraying as the number of black marks increases, Corollary 1 shows that if the random variable X is drawn from a distribution with a convex or linear distribution function (such as the uniform distribution), then the more black marks to date, the larger the marginal decrease in the likelihood of betraying. This follows from the convexity of $x^*(b)$. That is, under a convex distribution function, the likelihood of betraying decreases faster when the temptee's reputation consists of a larger number of betrayals. In other words, in this case the likelihood of rewarding increases faster when the temptee has a bad reputation.

The structure of the temptation game resembles settings where players have a choice between playing safe at some cost, or taking a risk of adding a "black mark." We briefly discuss the California criminal justice, driver's license suspension, tennis, baseball, and basketball below.

California criminal justice. California has a three-strikes-and-you-are-out rule[7] for criminals: one who gets convicted of three felonies gets jailed for life. In each period, a person can decide whether to commit a crime or not. If she commits a crime, there is the chance of being caught. Following our model, as she comes closer to getting put away for life (three convictions, hence three strikes), she is less likely to commit a crime. Consistent with our model, she could have a payoff from the crime if she does not get caught, her temptation, which might be the expected amount of money she would steal. In theory, recidivism rates in California should reveal a lesser propensity to criminal activity after two felony convictions. Unfortunately, two significant complications make this evidence almost impossible to assess. First, criminals are heterogeneous as to type. Those with two felony convictions presumably commit more crimes and are more likely to be caught, on average, than those with only one. Second, there is evidence that some two-strike criminals have migrated to other states.

Driver's license suspension. In some states, there are penalties for getting certain numbers of traffic infractions, or a certain number of accidents. If a driver gets a certain number of traffic violations in a period, then her license is suspended. Note that the motor vehicle bureau does not know how far a driver has traveled during that period, nor the level of temptation. (It may be more tempting to speed when one is late for an important meeting.) However, this is only part of the damage, since traffic infractions also cause a boost in insurance rates. This suggests that if we could fine people for a black mark in the temptation game, the additional instrument might afford a superior outcome.

Tennis. A tennis-player gets two serves. This is equivalent to being allowed two black marks before she is expelled (loses the point). On the first serve, it is optimal to be more aggressive. Indeed, the first serve is typically aggressive. It is struck with power and placement to have a high chance of winning the point outright or soon thereafter, assuming that it goes in. The second serve is usually much more conservative — slower speed, less risky placement — to make it exceedingly likely to go in and thereby avoid a double fault. This strategy has important, albeit not exact, parallels to being less willing to betray when one's reputation consists of more black marks in the temptation game. One key difference, of course, is that tennis is a game of strictly opposed interests, in

[7] The deterrent effect of strike laws has been studied in the legal literature (Shepherd, 2002; Helland and Tabarrok, 2007).

contrast to the temptation game. A second key difference is that a betrayal is a certain move, whereas whether a serve goes in is a probabilistic phenomenon.

Baseball. A related situation occurs in baseball. Holding the number of balls constant, batters can afford to be more picky on pitches when they have no strikes than with one strike, and with one strike than with two strikes.[8]

Basketball. In the NBA, if a player commits six personal fouls over the course of a game, he fouls out and is disqualified from participation for the remainder of the game. This is a setting, like our model, with imperfect monitoring. A player may get called for a foul that he did not commit, so it is dangerous to get up to five fouls. One of the features of basketball, but not the temptation game, is that an infraction is also punished by giving out foul shots. This situation is similar to the cumulative violations system in driving that was discussed above. Basketball players play less aggressively when they have more fouls, not wanting to take the chance of fouling out.

6. Optimal Cutoffs

This section identifies the pure equilibria (as characterized by Lemma 2) that are most favorable for the truster(s), most favorable for the temptees, and that optimize social welfare. We first discuss how those three might be chosen among all possible equilibria in practice.

We might think of these three situations as ones where the two different parties, or some uninvolved but benevolent coordinator, can select among the various possible equilibria. They may have this capability because they can establish customs or laws that apply in a particular community, or simply because they have the ability to communicate. Such communication can establish what Schelling (1980) labels a focal point. Myerson (2009)[9] comments on Schelling's insight: "anything in a game's environment or history that focuses the players' attention on one equilibrium may lead them to expect it, and so rationally to play it. This focal-point effect opens the door for cultural and environmental factors to influence rational economic behavior." Myerson goes on to observe that: "Schelling's focal-point effect should be counted as one of the most important ideas in social theory. Recognizing the fundamental problem of selecting among multiple equilibria can help us to better understand the economic impact of culture on basic social phenomena such as social relationships, property and justice, political authority and legitimacy, foundations of social institutions, reputations and commitment." Rosenthal and Landau (1979) study possible decision rules or "customs" which players might use to determine their moves in a game as a function of reputation. Our attention here is on ways of choosing among multiple equilibria in the temptation game.

[8] An opposite situation applies to the pitcher, of course. He must make more of an effort to get the ball over the plate when there are more balls thus far, holding strikes constant, lest he walk the batter.

[9] pages 1111-1112

In the temptation game, if only the truster(s) can communicate, they can say to the temptees: “We will allow you one betrayal, but once you get to two, no one will trust you.” If that message is transmitted, we would expect the temptees to behave as if $b^* = 1$. Cheap talk in such circumstances can determine which equilibrium is chosen: an equilibrium becomes focal because it is “agreed on” through cheap talk, and it is then followed (Farrell, 1987). See Crawford and Sobel (1982) and Farrell and Rabin (1996) for detailed discussions on cheap talk for coordination games and games of incomplete information. On the other hand, humans often use information on the opponent's past actions as coordinating devices, see Dalea et al. (2002) for an experimental study; in the context of the temptation game, if the truster(s) have not been trusting temptees with one black mark in the past, this may lead the temptees to expect that this behavior may continue in the future.

Often the players on one side actually have the power to establish customs or laws that can establish which equilibrium will prevail. For example, in most traditional societies men make the rules relating to marital behavior, such as the punishment for infidelity or the bases for divorce. Similarly, banks have historically made the rules for lending practices. But we could imagine a feminist movement changing the mores relating to marital behavior in a society[10] or, in the wake of the recent financial meltdown, we could easily envision Congress remaking the rules on lending to be more favorable to the consumer so as to maximize social welfare.

In the temptation game, players belong to identifiable groups, which may give them additional incentive to adhere to the rules set out in pre-game communications. If the players are to deviate from the announced group behavior, they will possibly suffer another type of reputation loss, with their peers.

6.1. An Inequality Between Optimal Cutoffs

In the temptation game, posit that Cs, and only Cs have the ability to communicate verbally. They will tell the Ds that they are employing the cutoff in number of black marks that maximizes the expected welfare of Cs assuming that Ds respond optimally to that cutoff. On the other hand, if the Ds have the power to choose the equilibrium — whether through communication or by setting the rules or laws in some group — they will commit to their x*(b) for $b \in \{0, 1, ..., b^*\}$, thus letting C pick his b^* in response; this produces the first best condition for D. Because C responds to D's strategy in a predictable way, D is effectively choosing C's cutoff b^* for him. The *socially optimal* equilibrium emerges when a third party, perhaps a government agency or an e-commerce site, proposes a set of strategies and associated equilibrium to optimize a weighted sum of the payoffs going to C and D.[11]

[10] For instance, in Lysistrata by Aristophanes, the women of Greece try to change the rules regarding who makes decisions about war. In particular, they withhold sexual privileges from their husbands as a means of forcing the men to negotiate peace.

[11] We note that the third party could also propose an equilibrium that achieves some other goal, e.g., if Amazon could specify the equilibrium that buyers and sellers play in the Amazon

We define $b^*(C)$ and $b^*(D)$ to be the optimal cutoffs for C and D respectively, that is, the cutoffs of the pure equilibria that maximize the corresponding payoffs. We let $b^*_\alpha(S)$ denote the cutoff that maximizes the sum of C's payoff and α times D's payoff, where $\alpha \geq 0$. If α is equal to 1, then $b^*_\alpha(S)$ maximizes the total payoff going to C and D. The following proposition shows that the optimal numbers of betrayals will be greatest for the temptee, moderate for the social optimum, and least for the truster.

Proposition 2. *For any $\alpha \geq 0$,*

$$b^*(C) \leq b^*_\alpha(S) \leq b^*(D) = B^*.$$

Recall that B^* is the maximum cutoff that can arise at a pure equilibrium (for given X, s, r and q). Proposition 2 tells us that the first best equilibrium for D has the maximum possible cutoff. Intuitively, D prefers to be trusted longer. Moreover, if the first best equilibrium for C has the maximum possible cutoff, then the equilibrium with $b^* = B^*$ is a social optimum. The first best equilibrium for C is the focus of the next two sections.

6.2. The Truster's Expected Payoff

Consider a pure equilibrium $(b^*, \{x^*(b), b = 0, ..., b^*\})$. Recall that $w(b) \equiv \mathbb{P}[X > x^*(b)]$ is the probability that player D betrays when she has b black marks. When C interacts with a D who has b black marks, his expected payoff is $1 - 2 \cdot w(b)$ if $b < b^*$, and 0 if $b = b^*$. In light of Proposition 1, this payoff increases in b when $b \in \{0, ..., b^* - 1\}$. A truster's payoff thus depends heavily on the number of black marks of the temptee (in the case that the truster interacts with one temptee per period) or the temptees (in the case that the truster interacts with a large number of temptees per period) that he interacts with in that period.

As far as C is concerned, the number of black marks of any D evolves according to a Markov chain. The state is the number of black marks, which either increases by 1 (if C trusts D and a betrayal is recorded), or remains the same (if either C trusts D and a reward is recorded or C does not trust D), or becomes 0 (if D dies and thus is replaced with a new player with a blank reputation).[12] Let π be the steady state distribution of this Markov chain. The following lemma gives C's expected payoff and provides a way to compute π.

Lemma 2. *Let $(b^*, \{x^*(b), b = 1, 2, ..., b^*\})$ be an equilibrium. The expected payoff of C from every D that he interacts with in a given period is equal to*

$$\sum_{b=0}^{b^*-1} \pi(b)(1 - 2w(b)), \tag{5}$$

Marketplace, it would perhaps choose the equilibrium that maximizes Amazon's revenue. We do not consider this situation in this paper.

[12] If D dies, she is replaced with a different D, but that does not affect C's payoff.

where $\pi(b)$ *satisfies*

$$\pi(0) = \frac{1-s}{1-s((1-q)(1-w(0))+rw(0))}; \tag{6}$$

and

$$\pi(b) = s\frac{q(1-w(b-1))+(1-r)w(b-1)}{1-s((1-q)(1-w(b))+rw(b))}\pi(b-1), \tag{7}$$

for $b \in \{1, ..., b^* - 1\}$.

In the next section, we demonstrate that $b^*(C)$ can involve the minimum (non-trivial) equilibrium cutoff, the maximum equilibrium cutoff, or an interior solution. C does not care how long a specific D lives, since he is guaranteed to meet a new D each period (in the setting of short-term interactions) or is interacting with a fixed large number of temptees in every period (in the setting of long-term interactions). However, he is interested in influencing the behavior of D, and the subtleties of that influence can yield these distinctive outcomes.

6.3. Numerical Examples

Proposition 2 shows that $b^*(C) \leq b^*(D)$. In this section we give some numerical examples to demonstrate that both $b^*(C) < b^*(D)$ and $b^*(C) = b^*(D)$ are possible. For simplicity, we consider perfect monitoring (i.e., $r = q = 0$) and assume that the random variable X has a uniform distribution.

Suppose X is uniform on $[\gamma, \delta]$ for some $\gamma < \delta$. Then

$$\mathbb{E}[(X - x^*)^+] = \frac{1}{\delta-\gamma}\int_{x^*}^{\delta}(x - x^*)dx = \frac{(\delta - x^*)^2}{2(\delta-\gamma)}.$$

Under perfect monitoring, (3) simplifies to

$$v(b) = v(b+1) + x^*(b)/s,$$

and (3) with (4) give

$$\frac{1-s}{s}x^*(b) + (1-s)v(b+1) = 1 + \frac{(\delta - x^*(b))^2}{2(\delta-\gamma)}.$$

The probability that player D betrays when she has b black marks is

$$w(b) = \frac{\delta - x^*(b)}{\delta - \gamma},$$

and the steady-state reputation distribution of the D population can be computed from the following equations (which follow from (6) and (7)).

$$\pi(0) = \frac{1-s}{1-s+sw(0)};$$

$$\pi(b) = s\frac{w(b-1)}{1-s-w(b)}\pi(b-1), b \in \{1, ..., b^* - 1\}.$$

We use these equations to compute the expected payoffs of C and D at the equilibria in the following examples.

Example 1. *Assume that X is uniform on $[0, 20]$ and $s = 0.95$. Then $B^* = 4$. C's payoff is maximized at $b^*(C) = 1$. That is, the one-betrayal-and-you-are-out strategy is best for C. This is the strategy that many societies have employed to deal with marital infidelities, particularly those of women. D's payoff on the other hand is maximized at $b^*(D) = 4$. Thus, in this case, $1 = b^*(C) < b^*(D) = 4$.*

Example 2. *Assume that X is uniform on $[0, 30]$ and $s = 0.95$. Then, $b^*(C) = 3$ and $b^*(D) = B^* = 4$.*

Example 3. *Assume that X is uniform on $[0, 1000]$ and $s = 0.95$. Then, for any $\alpha \geq 0$, $b^*(C) = b^*_\alpha(S) = b^*(D) = B^* = 3$, that is, both the truster and the temptee prefer the same cutoff.*

7. Adverse Selection

This paper has focused thus far on a pure moral hazard setting, where all temptees are the same in terms of payoff structure and self-control. Attention has thus been strictly on inducing good behavior despite temptation. This section allows for multiple types. Hence adverse selection rears its ugly head alongside moral hazard. Moreover, a truster will appropriately update his belief that a temptee is a particular type depending on the number of black marks she has received.

There are k types of temptees. Each type if defined by its distribution of temptations. Let X_i be the random variable that denotes the strength of the temptation to betray for type i. For a given cutoff b^*, we can compute the best response for type i (as in Section 4), which consists of a threshold $x^*_i(b)$ for each $b < b^*$. Each $x^*_i(b)$ is increasing and convex in b for $b \in \{0, ..., b^* - 1\}$ (this can be shown with the arguments used in the proof of Proposition 1). Moreover, the probability that a player of type i betrays is decreasing in b for $b \in \{0, ..., b^* - 1\}$. In words, the more black marks to date, the less likely a temptee of a particular type is to betray.

Which cutoffs b^* can arise in equilibrium in a world that allows for multiple types, hence adverse selection as well as moral hazard? A cutoff b^* can be an equilibrium if the truster's expected payoff from trusting is nonnegative for all $b \in \{0, 1, ..., b^* - 1\}$. Utilizing the framework of Section 4, we first compute the maximum cutoff that can arise at a (pure) equilibrium when the population is comprised solely of temptees of type i. Denote this maximum cutoff as B^*_i. Then, for any $b^* \leq \min_i B^*_i$, we have an equilibrium.[13]

The insights of Sections 5 and 6 still hold in a world with multiple types. Temptees still prefer the equilibrium with the maximum cutoff, whereas the

[13] Since $b^* \leq \min_i B^*_i$, we have that at each $b < b^*$ each type rewards with probability greater or equal to 0.5. This implies that a randomly chosen temptee with b black marks rewards with probability greater or equal to 0.5 when $b < b^*$. This in turn implies that the truster is better off trusting a temptee with $b < b^*$ black marks.

truster(s) may prefer a smaller cutoff. However, the truster's expected payoff is now given by a more complex formula than the one of Lemma 2, a formula that attends to the proportions of different types in the system.

In a world of multiple types, some will benefit from the presence of others, while others will lose. If the proportion of temptees of type j with $B_j^* = \min_i B_i^*$ is relatively small, there can be equilibria with $b^* > B_j^*$. Conversely, temptees of type h, where $B_h^* = \max_i B_i^*$, if they are not numerous, may find some of their forgiving (high b^*) equilibria disappear.

To gain additional insights, we specialize to a two-type world. Those types have temptations X_1 and X_2. Assume that X_2 stochastically dominates X_1, implying that the temptations to betray of type 2 are stronger than those of type 1. Accordingly, we refer to type 1 as the good type and type 2 as the bad type. The following lemma shows that at the maximum number of black marks at which the truster trusts the temptee, the bad type has a higher threshold x_i^* below which she rewards.[14]

Lemma 3. *If X_2 stochastically dominates X_1, then $x_1^*(b^* - 1) \leq x_2^*(b^* - 1)$.*

However, Lemma 3 says nothing about the probability that each type betrays with a particular black mark reputation, since the probability depends on the distribution. Interestingly, it is possible that the bad type's distribution of temptation stochastically dominates the good type's distribution of temptation, yet the bad type is less likely to betray at b black marks than is the good type. That is because she would be giving up more in terms of opportunity cost in the future. This will be the case if the threshold of the bad type $x_2^*(b)$ is sufficiently larger than the threshold of the good type $x_1^*(b)$ so that $F_1(x_1^*(b)) < F_2(x_2^*(b))$. We next provide an example where the good type is more likely to betray.

Example 4. *Suppose that*

$$X_1 = \begin{cases} 1 & \textit{with probability } 0.5 \\ 2 & \textit{with probability } 0.5 \end{cases}$$

$$X_2 = \begin{cases} 2 & \textit{with probability } 0.9 \\ 10 & \textit{with probability } 0.1 \end{cases}$$

We observe that X_2 stochastically dominates X_1.[15] *To avoid additional complication, assume that both types have the same survival probability $s = 0.6$ and that monitoring is perfect.*[16]

Suppose that $b^* = 1$. Then, at $b = 0$:

[14] We note that this may not hold for a smaller number of black marks. In particular, if $b < b^* - 1$, the threshold of the bad type may be lower than the threshold of the good type.

[15] In many examples, of course, neither type's distribution will stochastically dominate the other's.

[16] If survival probabilities differed, other factors equal, a type with better survival probabilities would be less likely to betray, since its members could better afford to wait for encounters where their temptations were high.

- *Type 1 rewards if $x = 1$ and betrays if $x = 2$. That is, a temptee of type 1 betrays with probability 0.5.*
- *Type 2 rewards if $x = 2$ and betrays if $x = 10$. That is, a temptee of type 2 betrays with probability 0.1.*

We conclude that in this case the good type is more likely to betray.

Multiple types introduce great richness to a world with black mark reputations. In particular, the distribution of types may shift over time with the number of black marks, a classic outcome with adverse selection. The shifting happens as the distribution of black marks converges to the steady state for every type of temptees. If initially all temptees have zero black marks and the number of temptees of each type is equal, initially a randomly selected temptee with zero black marks will belong to each type with probability 0.5. But as time goes by, a randomly selected temptee with zero black marks is more likely to be of the type that betrays with a smaller probability. For instance, in Example 4, in the steady state a temptee of type 1 has zero black marks with probability 0.57, whereas a temptee of type 2 has zero black marks with probability 0.87 (these numbers follow from (6)). Though 50% of those entering the system are of each type, individuals with 0 black marks will be 40% type 1s, whereas 77% of those with 1 black mark will be of type 1 once the system has converged to the steady state.

In the following variation of Example 4, type 1 is more likely (than type 2) to betray with no black marks, but less likely once a black mark is earned.

Example 5. *Suppose that*

$$X_1 = \begin{cases} 1 & \textit{with probability } 0.4 \\ 2 & \textit{with probability } 0.6 \end{cases}$$

$$X_2 = \begin{cases} 2 & \textit{with probability } 0.9 \\ 10 & \textit{with probability } 0.1 \end{cases}$$

Assume that both types have the same survival probability $s = 0.66$ and that monitoring is perfect.

Suppose that $b^* = 2$. At $b = 1$, type 1 is more *likely to betray than type 2: type 1 will betray whenever she has $X_1 = 1$ (i.e., with probability 0.4), whereas type 2 will betray only if she has $X_2 = 10$ (i.e., with probability 0.1). On the other hand, at $b = 0$, type 1 is* less *likely to betray than type 2: type 1 betrays with probability 0.4, whereas type 2 betrays with probability 1.*

In real world play, there are sure to be multiple types with temptations that vary over time. Temptees will exhibit both adverse selection and moral hazard. Trusters will seek to deter their bad behavior with an ultimate refusal to play once black marks reach a certain level.

8. Mechanisms that also Track the Number of Transactions

Thus far, we have only considered mechanisms that track the number of recorded betrayals. This section extends the reputation mechanism to also include information on the *number of transactions*, that is, the number of times that the temptee has been trusted. To keep matters simple, we deal with only a single type of temptee. We show that if the number of transactions is recorded, the maximum number of black marks that C allows D at a pure equilibrium does not increase compared to the case where the number of transactions is not recorded. In particular, at a pure equilibrium, C will never trust D once she has B^* black marks — regardless of whether C knows the number of transactions of D.

We assume that D's reputation consists of the number of black marks b and the number of transactions (which we denote by n) that she has completed. We thus denote D's reputation by (b, n). We refer to this as the two-dimensional case, in contrast to the one-dimensional case where the temptee's reputation consists only of the number of black marks. A strategy of player C in this more general model consists of a cutoff for each number of transactions. With a slight abuse of notation, let $b^*(n)$ be the cutoff for n transactions, that is, C does not trust D at (b, n) if $b \geq b^*(n)$. On the other hand, D's strategy will consist of a threshold $x^*(b, n)$ for every possible reputation (b, n). That is, when D has b black marks in n transactions, then she betrays if the strength of her temptation to betray (i.e., the realization of X) exceeds the threshold $x^*(b, n)$.

The sets $\{b^*(n), n = 0, 1, ...\}$ and $\{x^*(b, n), b = 0, 1, ..., b^*(n), n = 0, 1, ...\}$ constitute a pure equilibrium of the temptation game if the following conditions are satisfied:

1. $x^*(b, n)$ is a best response of D to C's strategy, i.e., there exists a function $v(b, n)$ such that for $b < b^*(n)$:

$$v(b,n) = 1+s\cdot(1-q)\cdot v(b,n+1)+s\cdot q\cdot v(b+1,n+1)+\mathbb{E}[(X-x^*(b,n))^+]; \quad (8)$$

$$x^*(b,n) = s \cdot (1-r-q) \cdot (v(b,n+1) - v(b+1,n+1)); \quad (9)$$

and

$$v(b^*(n),n) = 0 \text{ for all } n. \quad (10)$$

2. $b^*(n)$ is a best response of C to D's strategy, i.e.,
   - $\mathbb{P}[X < x^*(b,n)] \geq 1/2$ for $b < b^*(n)$
   - $\mathbb{P}[X < x^*(b^*,n)] \leq 1/2$

The derivation of these equilibrium conditions is similar to the derivation for the one-dimensional case in Section 4. Moreover, this set of conditions is a generalization of conditions of Section 4. In particular, if we set $b^*(n) = b^*$ where $b^* \leq B^*$, then we get an equilibrium of the two-dimensional reputation mechanism, which corresponds to an equilibrium of the one-dimensional mechanism; in this case, $b^*(n)$ is a constant that does not vary with the number of transactions n. However, in the two-dimensional case, there also exist equilibria where $b^*(n)$ varies with n.

We next show two fundamental properties of $b^*(n)$. First, $b^*(n)$ is non-decreasing in n. This is an intuitive property: the larger the number of transactions of D, the larger the number of black marks that C will tolerate. Second, $b^*(n)$ is bounded above by B^*, i.e., the maximum cutoff for which there exists a pure equilibrium when reputation only consists of the number of black marks. Even though the number of transactions is recorded in the two-dimensional reputation case, C does not allow D to commit more betrayals than in the one-dimensional case.

Proposition 3. *If reputation consists of both the number of black marks b and the number of transactions n, then at any pure equilibrium of the temptation game*

(i) $b^*(n)$ *is non-decreasing*
(ii) $b^*(n) \leq B^*$ *for all* n

The fact that $b^*(n)$ is upper-bounded by B^* may at first seem counterintuitive. One could expect that C would allow D more black marks when he knows that she has completed a very large number of transactions than when he has no information on the number of transactions. However, if C tolerated a large number of transactions, then D would betray with high probability, since one more betrayal would make a small difference in her future expected payments. This would be similar to a reputation mechanism that aggregates ratings over the lifetime of the temptee; an approach that has been shown to be ineffective in a number of settings (e.g., Fan et al., 2005; Aperjis and Johari, 2010; Cripps et al., 2004).

We get some additional intuition for why $b^*(n)$ is upper-bounded by B^* by considering the steps of the proof. We first observe that $b^*(n)$ cannot be very large; if it were, D would not be properly incentivized at $(0, n)$. In particular, even though betraying at $(0, n)$ would bring her closer to expulsion, expulsion would be so far in the future that D would not reward with a sufficiently large probability (i.e., she would use a small $x^*(b, n)$), and C on his side would not trust her. Now, given that $b^*(n)$ is upper bounded and increasing, we conclude that $b^*(n)$ is constant for all large n, which means that the number of transactions does not play any role after some point. Then, after a sufficiently large number of transactions, this two-dimensional problem is equivalent to the setting where only the number of black marks is included in D's reputation. As a result, the number of black marks that C tolerates does not exceed B^*.

As mentioned above, after a certain number of transactions $b^*(n)$ becomes constant. Then, the thresholds $x^*(b, n)$ correspond to thresholds of a one-dimensional equilibrium. Thus, after that point D is less likely to betray when she has more black marks ($x^*(b, n)$ is increasing in b). However, $x^*(b, n)$ may not be increasing in b for small values of n when there is a high probability of misrecording a betrayal.[17]

9. Conclusion

This paper studies how reputations work when the truster's decision to trust is based only on the number of recorded betrayals of the temptee (black mark reputation mechanism). We find that at a pure equilibrium, the greater the number of black marks, the less likely the temptee is to betray. This insight applies to a broad range of situations where black marks are recorded, such as agencies for consumer protection and online review sites.

Black mark reputations approximate settings where buyers focus on the number of negatives — even if more reputation information is provided. The Availability Heuristic (Tversky and Kahnenman, 1973) leads individuals to judge the frequency of an event by how easily they can bring an instance to mind. This heuristic leads individuals to give significant weight to extreme bad outcomes. Recognizing this, we have been told that in the venture capital industry, executives work extremely hard to prevent portfolio companies from going bankrupt, even though that same time devoted to profitable ventures would yield greater benefits to both the executives and the investors. The VC executives recognize that their world is an approximation of the black marks world, where embarrassing strikeouts are remembered, and batting averages are sometimes forgotten.

We recognize that to accurately describe some interactions between trusters and temptees, the model of this paper would have to be elaborated. We show in Section 8 that the same qualitative results apply in the long run when the number of transactions is also recorded as part of a temptee's reputation. Two further extensions suggest themselves immediately. First, some relationships have a natural termination or sunset date quite apart from black marks. Thus, for a college and a student, rules infractions — plagiarism or disorderly behavior — would be the equivalent of betrayals. But once graduation occurs, the relationship is ended no matter what; past black marks become irrelevant. Second, many long-term relationships — and some one-time-only relationships — have both parties trusting and both parties tempted. Thus, the business and its supplier or the husband and the wife may both rely on each other; each has a reputation and each can betray.

Across a wide swath of societal concerns, we live with the notion that a single betrayal does not end a relationship. Thus, there are second chances (and possibly more). Religions routinely allow for forgiveness. "The God I believe in is a God of second chances," Bill Clinton once said, referring to his own shortcomings. And George W. Bush, not known for being soft on crime, observed: "America is the land of the second chance — and when the gates of the prison open, the path ahead should lead to a better life." That is the way two successive Presidents outlined the theme that motivated this analysis:

[17] We thank John H. Lindsey II for constructing an example where $x^*(b, n) > x^*(b+1, n)$ and $b + 1 < b^*(n)$ at an equilibrium.

The game of life accommodates betrayals, but not without putting betrayers on warning.

From the time of the snake in the Garden of Eden, temptation has always been with us. Benjamin Franklin, a prolific provider of adages in a much earlier era, once wrote: “What makes resisting temptation difficult for many people, is that they don’t want to discourage it completely.” Betrayals must be expected from all of us, and reputations are required to keep them within bounds. And should betrayals exceed some critical value, expulsion will be our fate. Such is the life of the temptee.

Appendix A: Proofs

Proof of Lemma 1: Consider some $\beta \in \{1, ..., b^* - 1\}$. The optimal strategy of D when starting at β is $\{x^*(b), b = \beta, ..., b^* - 1\}$. We now consider $\beta - 1$, and assume that D uses the following strategy: at $\beta - 1$ she rewards if $X < x^*(\beta)$, at β she rewards if $X < x^*(\beta + 1)$,..., at $b^* - 1$ she rewards if $X < x^*(b^* - 2)$, and at $b^* - 1$ she rewards for any X. That is, she uses the optimal strategy for β starting at $\beta - 1$ and then always rewards if her reputation consists of $b^* - 1$ betrayals. Up to (and including) state $b^* - 2$, this yields payoff $v(\beta)$ since everything is the same as when starting at β. Then, D reaches state $b^* - 1$ with positive probability and gets a strictly positive payment once she is there. Therefore, when starting at $\beta - 1$ the described strategy yields a strictly higher payoff than $v(\beta)$; thus $v(\beta - 1) > v(\beta)$. □

Proof of Proposition 1: We first show that $x^*(b)$ is strictly increasing in b for $b \in \{0, ..., b^* - 1\}$. (3) can be rewritten as

$$v(b) = \frac{x^*(b)}{s(1 - r - q)} + v(b + 1)$$

Substituting in (4) we have

$$\frac{1 - s(1 - q)}{s(1 - r - q)} x^*(b) + (1 - s)v(b + 1) = 1 + \mathbb{E}[(X - x^*(b))^+]. \tag{11}$$

Let $b_1 < b_2$, and let $x_1 = x^*(b_1)$ and $x_2 = x^*(b_2)$ be the corresponding solutions of (11). Then,

$$\frac{1 - s(1 - q)}{s(1 - r - q)} x_1 + (1 - s)v(b_1 + 1) = 1 + \mathbb{E}[(X - x_1)^+]. \tag{12}$$

$$\frac{1 - s(1 - q)}{s(1 - r - q)} x_2 + (1 - s)v(b_2 + 1) = 1 + \mathbb{E}[(X - x_2)^+]. \tag{13}$$

Suppose that $x_1 \geq x_2$. Then,

$$\begin{aligned} &\frac{1 - s(1 - q)}{s(1 - r - q)} x_2 + (1 - s)v(b_2 + 1) < \\ &\frac{1 - s(1 - q)}{s(1 - r - q)} x_1 + (1 - s)v(b_1 + 1) = \\ &1 + \mathbb{E}[(X - x_1)^+] \leq \\ &1 + \mathbb{E}[(X - x_2)^+], \end{aligned}$$

which contradicts (13). We note that the first inequality follows because v is decreasing in b (by Lemma 1) and $s < 1$; the equality follows from (12), and the second inequality holds because $x_1 \geq x_2$. We conclude that $x_1 < x_2$, and thus $x^*(b)$ is strictly increasing in b for $b \in \{0, 1, ..., b^* - 1\}$.

We next show that $x^*(b)$ is convex in b for $b \in \{0, ..., b^* - 1\}$. Let

$$g(y) \equiv 1 + \mathbb{E}[(X - y)^+] = 1 + \int_y^\infty (x - y) dF(x).$$

We observe that $g'(y) = -(1 - F(y))$. This implies that $g'(y)$ is negative and increasing in y, and thus g is decreasing and convex. From (11) we find that

$$\frac{1 - s(1-q)}{s(1-r-q)}(x^*(b) - x^*(b-1)) + (g(x^*(b-1)) - g(x^*(b))) = (1-s)(v(b) - v(b+1))$$

Moreover, by (3) we have that $v(b) - v(b+1) = x^*(b)/(s(1-r-q))$. Thus,

$$\frac{1 - s(1-q)}{s(1-r-q)}(x^*(b) - x^*(b-1)) + (g(x^*(b-1)) - g(x^*(b))) = \frac{1-s}{s(1-r-q)} x^*(b).$$

Let $b_1 < b_2$. Since $x^*(b)$ is increasing in b (by the first part of this proof), we have that

$$\begin{aligned} &\frac{1 - s(1-q)}{s(1-r-q)}(x^*(b_1) - x^*(b_1 - 1)) + (g(x^*(b_1 - 1)) - g(x^*(b_1))) < \\ &\frac{1 - s(1-q)}{s(1-r-q)}(x^*(b_2) - x^*(b_2 - 1)) + (g(x^*(b_2 - 1)) - g(x^*(b_2))) \end{aligned} \tag{14}$$

Suppose that $x^*(b_1) - x^*(b_1 - 1) > x^*(b_2) - x^*(b_2 - 1)$. Then, by the convexity of g we have that

$$\begin{aligned} &g(x^*(b_1 - 1)) - g(x^*(b_1)) \geq \\ &g(x^*(b_2 - 1)) - g(x^*(b_2 - 1) + (x^*(b_1) - x^*(b_1 - 1))) \geq \\ &g(x^*(b_2 - 1)) - g(x^*(b_2 - 1) + (x^*(b_2) - x^*(b_2 - 1))) \geq \\ &g(x^*(b_2 - 1)) - g(x^*(b_2)) \end{aligned}$$

which contradicts (14). We note that the first inequality holds because g is convex, the second inequality is a consequence of $x^*(b_1) - x^*(b_1 - 1) > x^*(b_2) - x^*(b_2 - 1)$, and the third inequality holds because g is decreasing. Thus, $x^*(b) - x^*(b-1)$ is nondecreasing in b and $x^*(b)$ is convex for $b \in \{0, 1, ..., b^* - 1\}$. □

Proof of Proposition 2: We first show that $b^*(D) = B^*$. Let $u(b, b^*)$ be equal to $v(b)$ when the cutoff b^* is used. We observe that $u(b, b^*)$ only depends on the difference $b^* - b$ (given the same s, r and q), and is increasing in $b^* - b$. Thus, $u(0, b^*)$ is maximized when b^* is maximized.

Now that we have shown that $b^*(C) \leq b^*(D)$, how about the socially optimal equilibrium $b^*_\alpha(S)$ which optimizes the weighted return of C and D? Because the return for D decreases when b^* decreases, it is not possible that $b^*_\alpha(S) < b^*(C)$, because both players would be better off with $b^*_\alpha(S) = b^*(C)$. It is also not possible that $b^*_\alpha(S) > b^*(D)$, because $b^*(D)$ is the highest b^* possible in the equilibrium set. Then we have $b^*(C) \leq b^*_\alpha(S) \leq b^*(D)$ for all $\alpha \geq 0$. □

Proof of Lemma 2: Let $\pi(i)$ be the probability that a randomly chosen D player has i betrayals. For $i = 0$ the balance equation is:

$$\pi(0) = s((1-q)(1-w(0)) + rw(0))\pi(0) + (1-s).$$

In particular, when D has no betrayals, she rewards with probability $1 - w(0)$ and betrays with probability $w(0)$. If she rewards, her reputation remains at $b = 0$ with probability $1-q$; if she betrays, her reputation remains at $b = 0$ with probability r. Finally, from any state, there is a transition to 0 with probability $1 - s$.

Similarly, for $b \in \{1, ..., b^* - 1\}$ we have

$$\pi(b) = s(q(1-w(b-1))+(1-r)w(b-1))\pi(b-1)+s((1-q)(1-w(b))+r\cdot w(b))\pi(b)$$

Solving for $\pi(0)$ and $\pi(b)$ we get (6) and (7) respectively. □

Proof of Lemma 3: For $b = b^* - 1$, (3) can be rewritten as

$$v_i(b^* - 1) = \frac{x_i^*(b^* - 1)}{s(1-r-q)}$$

for each type i, because $v_i(b^*) = 0$. Substituting in (4) we have

$$\frac{1-s(1-q)}{s(1-r-q)} x_1^*(b^* - 1) = 1 + \mathbb{E}[(X_1 - x_1^*(b^* - 1))^+];$$

$$\frac{1-s(1-q)}{s(1-r-q)} x_2^*(b^* - 1) = 1 + \mathbb{E}[(X_2 - x_2^*(b^* - 1))^+].$$

Suppose for the sake of contradiction that $x_1^*(b^* - 1) > x_2^*(b^* - 1)$. Then, it must be that

$$\mathbb{E}[(X_1 - x_1^*(b^* - 1))^+] > \mathbb{E}[(X_2 - x_2^*(b^* - 1))^+].$$

However, this cannot hold if X_2 stochastically dominates X_1 and $x_1^*(b^* - 1) > x_2^*(b^*-1)$ (because then $X_2 - x_2^*(b^*-1)$ stochastically dominates $X_1 - x_1^*(b^*-1)$). We conclude that $x_1^*(b^* - 1) \leq x_2^*(b^* - 1)$. □

Proof of Proposition 3: Suppose that $b^*(n) > b^*(n+1)$. Suppose that D has $b^*(n+1)$ betrayals and n transactions. C will trust D in this state, because $b^*(n) > b^*(n+1)$. However, D knows that whatever she does in this period, C will not trust her in the next period. So then, it is optimal for her to betray. But if D betrays, C has no reason to trust. So it must be that $b^*(n) \leq b^*(n+1)$, which shows (i).

To show (ii), we first show that at a pure equilibrium,

$$b^*(n) \leq \frac{\log((1+\mathbb{E}X)/m) + \log(1/(1-s))}{\log(1/s)} \equiv K.$$

Consider an equilibrium where C's strategy is given by $\{b^*(n)\}$ and D's strategy is $\{x^*(b,n)\}$ and suppose that there exists some n with $b^*(n) > K$. A necessary condition for this to be an equilibrium is that $x^*(0,n) \geq m$. We show that D is better off using some strategy $\{x(b,n)\}$ with $x(0,n) < m$. In particular, consider a strategy with $x(b,n') = x^*(b,n')$ for $n' > n$. Suppose that D's current reputation is $(0,n)$ and let x be the current realization of X. If D betrays now, her current payment will increase by x. We next upper bound the amount that D will lose by betraying now. First observe that the earlier time that D may be expelled is K periods later. This is a very conservative estimate, because D will probably not always betray and $b^*(n)$ may be increasing. We next observe that D misses at most $1 + \mathbb{E}X$ in expectation for each period after she is expelled. Considering that D only survives with probability s in every period, in total she misses at most $(1 + \mathbb{E}X)/(1-s)$. But this payment is at least K periods away, so she discounts it by at most s^K. Thus, if D betrays now, her future payment will decrease by at most

$$\frac{s^K}{1-s}(1 + \mathbb{E}X).$$

We conclude that D will be better off betraying if

$$x > \frac{s^K}{1-s}(1 + \mathbb{E}X).$$

Because of the way we defined K, note that

$$\frac{s^K}{1-s}(1 + \mathbb{E}X) < m.$$

So D is better off using $x(0,n) < m$.

Thus, for any given problem there exists some constant upper bound on $b^*(n)$. We already know that $b^*(n)$ is non-decreasing, so there must exist a $\bar{n}$ and a $\bar{b}$ such that $b^*(n) = \bar{b}$ for $n \geq \bar{n}$. Then, after $\bar{n}$ the exact number of transactions does not affect the cutoff (which is constant). Without loss of generality, we can restrict the state-space to

$$\{(b,n) : b \leq b^*(n), n < \bar{n}\} \cup \{(b,\bar{n}) : b \leq \bar{b}\},$$

which is finite.

For $n < \bar{n}$, equations (8), (9), and (10) need to be satisfied. For $n = \bar{n}$, we have

$$v(b,\bar{n}) = 1 + s \cdot (1-q) \cdot v(b,\bar{n}) + s \cdot q \cdot v(b+1,\bar{n}) + \mathbb{E}[(X - x^*(b,\bar{n}))^+] \text{ for } b < \bar{b};$$

$$x^*(b,\bar{n}) = s(1-r-q) \cdot (v(b,\bar{n}) - v(b+1,\bar{n})) \text{ for } b < \bar{b};$$

$$v(\bar{b},\bar{n}) = 0.$$

Observe that $\bar{n}$ is just a dummy variable in the previous equations. More importantly, if we ignore $\bar{n}$ these are exactly equations (1), (3), and (4), that is, the equations we had in the one-dimensional case, where D's reputation consisted only of the number of betrayals. This observation implies that $\bar{b} \leq B^*$, and also $b^*(n) \leq B^*$, which concludes the proof for (ii). □

Appendix B: Mixed-Strategy Equilibria

Our analysis in this paper considers pure-strategy equilibria. For completeness, we now discuss mixed strategy equilibria in the same framework. In this appendix, we consider mixed strategies for player C, where C trusts D with some probability that depends on her reputation. We consider the same model as in section 3 and allow for imperfect monitoring, assuming that *D's reputation does not change in periods that she did not interact with a C because she was not trusted by the C she was matched to.*

We consider settings where C may use a mixed strategy: C's mixed strategy consists of the *probability* he trusts D as a function of her reputation. This is a generalization of the pure strategy where C trusts D with either probability 1 or probability 0. We do not consider mixed strategies for player D here, and assume that X is continuous for simplicity.[18] Thus, D's strategy shows (as in the pure equilibrium case) whether she rewards as a function of her reputation.

We start by characterizing mixed-strategy equilibria. We then consider a special class of mixed equilibria, at which C trusts D strictly more than at any pure equilibrium. We show that the temptee strictly prefers such an equilibrium to a pure equilibrium. On the other hand, we conjecture that the truster is better off at a pure equilibrium.

Characterization of Mixed-Strategy Equilibria

C's mixed strategy represents the *probability* he trusts D as a function of her reputation. Thus, C's strategy is summarized by $\{p^*(b), b = 0, 1, ...\}$, where $p^*(b)$ is the probability that C trusts D when her reputation consists of b black marks. On the other hand, we do not consider mixed strategies for player D; thus the set $\{x^*(b), b = 0, 1, ...\}$ represents D's strategy (as in the pure equilibrium case).

As before, let $v(b)$ be the maximum expected infinite horizon payoff to D when she has b betrayals. Then,

$$\begin{aligned} v(b) = (1 - p^*(b))sv(b) + p^*(b)\,(1 + \mathbb{E}[\max\{&X + s((1-r)v(b+1) + r \cdot v(b)), \\ &s((1-q)v(b) + q \cdot v(b+1))\}]) \end{aligned}$$

In particular, with probability $p^*(b)$, C trusts D and then D chooses whether to reward or betray. On the other hand, with probability $1 - p^*(b)$, C does not trust D. In that case, D receives zero payment in this period and her reputation

[18] The results can be easily extended to consider mixed strategies for player D; such strategies would only be relevant if X follows a discrete distribution. In particular, D would only randomize if $X = x^*(b)$. If X is drawn from a continuous distribution (as we are assuming in this appendix), X will be equal to $x^*(b)$ with zero probability, and thus what D does at $x^*(b)$ does not affect the players' payoffs. If X is drawn from a discrete distribution, she could mix optimally at most at one point (i.e., $x^*(b)$).

remains the same. Straightforward calculations show that

$$\begin{aligned} v(b) =& p^*(b)\left(1 + s(1-q)v(b) + s \cdot q \cdot v(b+1) + \mathbb{E}[(X - x^*(b))^+]\right) \\ &+ (1-p^*(b))s \cdot v(b), \end{aligned}$$

where

$$x^*(b) \equiv s \cdot (1-r-q) \cdot (v(b) - v(b+1))$$

On the other hand, C trusts D if $\mathbb{P}[X > x^*(b)] > 1/2$; C does not trust D if $\mathbb{P}[X > x^*(b)] < 1/2$; and C is indifferent between trusting and not trusting if $\mathbb{P}[X > x^*(b)] = 1/2$. Thus, if $x^*(b) > m$, then C plays $p^*(b) = 1$; if $x^*(b) = m$, any $p^*(b) \in [0,1]$ is a best response for C; and if $x^*(b) < m$, then C plays $p^*(b) = 0$ (where m is the median of X).

We conclude that the sets $\{x^*(b), b = 0, 1, ...\}$, $\{p^*(b), b = 0, 1, ...\}$ constitute an equilibrium if the following conditions are satisfied:

1. $\{x^*(b), b = 0, 1, ...\}$ is a best response of D to C's strategy, i.e., there exists a function $v(b)$ such that

$$x^*(b) = s \cdot (1-r-q) \cdot (v(b) - v(b+1)); \tag{15}$$

$$(1 - s(1 - q \cdot p^*(b)))v(b) = p^*(b)\left(1 + s \cdot q \cdot v(b+1) + \mathbb{E}[(X - x^*(b))^+]\right). \tag{16}$$

2. $p^*(b)$ is a best response of C to D's strategy, i.e.,
   - if $\mathbb{P}[X < x^*(b)] > 1/2$ then $p^*(b) = 1$
   - if $\mathbb{P}[X < x^*(b)] = 1/2$ then $p^*(b) \in [0,1]$
   - if $\mathbb{P}[X < x^*(b)] < 1/2$ then $p^*(b) = 0$.

Pure equilibria are a special case of the mixed equilibria discussed here. In particular, if $p^*(b) = 1$, then Equation (16) yields (4). We note that Equation (15) is the same as (3), rewritten here for convenience.

Given an equilibrium $\{(x^*(b), p^*(b)), b = 0, 1, ...\}$, we define

$$b^* \equiv \min\{b : p^*(b) = 0\}.$$

In words, b^* is the cutoff (in terms of the number of black marks) at which C does not trust D. If D is never trusted when her reputation consists of b^* black marks, then she never has the chance to reach more than b^* black marks.

We already know that a finite cutoff (after which C stops trusting D) is associated with every pure equilibrium. The following lemma shows that this is also the case for mixed equilibria.

Lemma 4. *For every equilibrium $\{(x^*(b), p^*(b)), b = 0, 1, ...\}$ there exists a finite cutoff b^* such that $p^*(b) > 0$ for $b < b^*$ and $p^*(b^*) = 0$.*

Proof. Suppose that $p^*(b) > 0$ for all b. Then, (15) implies that

$$v(0) = \frac{1}{s(1-r-q)} \sum_{b=0}^{\infty} x^*(b) \geq \frac{1}{s(1-r-q)} \sum_{b=0}^{\infty} m = \infty,$$

because $m > 0$.

On the other hand, by substituting (15) in (16), we have that

$$(1-s)v(0) = p^*(0)\left(1 + \mathbb{E}[(X - x^*(b))^+] - \frac{q}{1-r-q}x^*(b)\right),$$

which cannot hold if $v(0) = \infty$, since $\mathbb{E}X < \infty$ and $p^*(0) \leq 1$. We conclude that there always exists a finite cutoff. □

Since in this section we are assuming that X is continuous, the condition $\mathbb{P}[X < x^*(b)] = 1/2$ is equivalent to $x^*(b) = m$.

Proposition 1 shows that for pure equilibria, $x^*(b)$ is strictly increasing and convex for b in $\{0, 1, ..., b^* - 1\}$. This is not the case for mixed equilibria in general, since at a mixed equilibrium we may have $x^*(b^*-1) = m$ and $x^*(b) > m$ for $b < b^* - 1$. However, the insights of Proposition 1 still hold if we do not consider the b's for which $x^*(b) = m$; this is discussed in Example 6.

We next show how C's expected payoff is computed in a mixed equilibrium, by providing a generalization of Lemma 2.

Lemma 5. *Let $(b^*, \{p^*(b), b = 1, 2, ..., b^*\}, \{x^*(b), b = 1, 2, ..., b^*\})$ be an equilibrium. The expected payoff of C is equal to*

$$\sum_{b:p^*(b)=1} \pi(b)(1 - 2w(b)),$$

where $\pi(b)$ satisfies

$$\pi(0) = \frac{1-s}{1 - s((1-q)(1-w(0)) + rw(0))p^*(0) - (1 - p^*(0))};$$

and

$$\pi(b) = s\frac{q \cdot (1 - w(b-1)) + (1-r)w(b-1)}{1 - s((1-q)(1-w(b)) + rw(b))p^*(b) - (1 - p^*(b))}p^*(b-1)\pi(b-1)$$

for $b \in \{1, ..., b^* - 1\}$.

Proof. Let $\pi(b)$ be the probability that a randomly chosen D player has b betrayals. For $b = 0$ the balance equation is:

$$\pi(0) = s(((1-q)(1-w(0)) + rw(0))p^*(0) + (1 - p^*(0)))\pi(0) + (1-s).$$

In particular, when D has no betrayals, C trusts her with probability $p^*(0)$. If D is trusted, then she rewards with probability $1 - w(0)$ and betrays with probability $w(0)$. If she rewards, her reputation remains at $b = 0$ with probability $1 - q$; if she betrays, her reputation remains at $b = 0$ with probability r. Finally, from any state, there is a transition to 0 with probability $1 - s$.

Similarly, for $b \in \{1, ..., b^* - 1\}$ we have

$$\begin{aligned}\pi(b) =& s(q(1 - w(b-1)) + (1-r)w(b-1))p^*(b-1)\pi(b-1) \\ &+ s(((1-q)(1-w(b)) + rw(b))p^*(b) + (1 - p^*(b)))\pi(b)\end{aligned}$$

Solving for $\pi(0)$ and $\pi(b)$, we get the equations of the lemma.

C's expected payoff is then equal to

$$\sum_{b=0}^{b^*-1} p^*(b)\pi(b)(1-2w(b))$$

If $p^*(b)$ is strictly between 0 and 1, then $x^*(b) = m$ and $w(b) = 1/2$, and thus C's expected payoff in that period is equal to 0. This observation implies that C's expected payoff can be equivalently written as

$$\sum_{b:p^*(b)=1} \pi(b)(1-2w(b)).$$

□

Dominant Extend Equilibria

In this section we study a particular category of mixed equilibria that (when they exist) prolong trust compared to pure equilibria. Recall that B^* is the maximum cutoff that arises in a *pure* equilibrium. Then, at the best possible pure equilibrium for D, C trusts her until she has B^* black marks (by Proposition 2). Often, there exist mixed equilibria at which C always trusts D until she has B^* black marks and then at B^* black marks he trusts her with some probability that is strictly between 0 and 1. We call such equilibria *dominant extend equilibria.* A precise definition is given below.

Definition 1. *A dominant extend equilibrium satisfies*

- $p^*(b) = 1$ *for* $b = 0, 1, ..., B^* - 1$
- $p^*(B^*) \in (0, 1)$
- $p^*(B^* + 1) = 0$

A dominant extend equilibrium gives D a longer expected lifetime, and, as the following proposition shows, D always prefers a dominant extend equilibrium to any pure equilibrium.

Proposition 4. *D strictly prefers a dominant extend equilibrium to a pure equilibrium.*

Proof. We show that D's payoff is strictly greater at a dominant extend equilibrium. Consider the pure equilibrium with $b^* = B^*$, and suppose that D's strategy is $\{x^*(b), b = 0, 1, ..., b^* - 1\}$. Let $v_{PE}(0)$ be D's expected payoff at this pure equilibrium. Now suppose there exists a dominant extend equilibrium with cutoff $b^* = B^* + 1$. D's expected payoff in the dominant extend equilibrium will be at least as much as the payoff she would get by using $x^*(b)$ for $b = 0, 1, ..., B^* - 1$ and always rewarding at B^*. This strategy yields payoff $v_{PE}(0)$ up to state $B^* - 1$. Then, D reaches a reputation of B^* betrayals with positive probability; and once she has B^* betrayals, C trusts her with positive probability. Thus, in expectation she gets a strictly positive payoff once she has B^* betrayals. Therefore, D's payoff at the dominant extend equilibrium is strictly greater than her payoff at a pure equilibrium. □

Dominant extend equilibria are always attractive to D, because D's expected payoff is maximized the longer she can expect to live. Thus, rather than being expelled for sure at $b = B^*$, she would prefer to have a probabilistic chance there, with expulsion at $b = B^* + 1$.

On the other hand, we conjecture that player C always prefers a pure equilibrium to a dominant extend equilibrium. Our intuition for this is as follows. Consider the pure equilibrium with the maximum cutoff and the dominant extend equilibrium. We use the subscripts PE and DEE to denote the two equilibria. By Lemma 5 it suffices to show that

$$\sum_{b=0}^{B^*-1} \pi_{PE}(b)(1 - w_{PE}(b)) \geq \sum_{b=0}^{B^*-1} \pi_{DEE}(b)(1 - w_{DEE}(b)). \tag{17}$$

Lemma 6 (below) shows that $w_{PE}(b) \leq w_{DEE}(b)$. On the other hand, the number of states is larger at the dominant extend equilibrium, suggesting that (17) is plausible. We conducted extensive simulations that support this claim.

Lemma 6. *For $b \in \{0, 1, ..., B^* - 1\}$, $w_{PE}(b) \leq w_{DEE}(b)$.*

Proof. It suffices to show that $x^*_{PE}(i) \geq x^*_{DEE}(i)$, since w is decreasing in x^*.

We rewrite (11) from the proof of Proposition 1.

$$\frac{1 - s(1-q)}{s(1-r-q)} x^*(b) + (1-s)v(b+1) = 1 + \mathbb{E}[(X - x^*(b))^+].$$

Since the left-hand side is increasing in $x^*(b)$ and the right-hand side is decreasing in $x^*(b)$, the solution $x^*(b)$ decreases as $v(b+1)$ increases. (The formal proof of this is similar to the first part of the proof of Proposition 1.) We use this fact to show that $x^*_{PE}(i) > x^*_{DEE}(i)$ for $b \in \{0, 1, ..., B^* - 1\}$ by induction. The induction hypothesis is that $x^*_{PE}(i) \geq x^*_{DEE}(i)$ and $v_{PE}(i + 1) < v_{DEE}(i + 1)$.

- Basis: $i = B^* - 1$. We have that $v_{PE}(B^*) = 0$, while $v_{DEE}(B^*) > 0$. Thus $v_{PE}(B^*) < v_{DEE}(B^*)$ and $x^*_{PE}(B^* - 1) > x^*_{DEE}(B^* - 1)$.
- Induction Step: Suppose that $x^*_{PE}(i) > x^*_{SE}(i)$ and $v_{PE}(i+1) < v_{DEE}(i+1)$. Then, by (4), $v_{PE}(i) < v_{DEE}(i)$, which in turn implies that $x^*_{PE}(i-1) > x^*_{DEE}(i-1)$ and completes the induction step.

□

If player C always prefers a pure equilibrium, and C is the one that specifies which equilibrium will be played, then we will have a pure equilibrium. On the other hand, pure equilibria often tend to arise as focal points because of their simplicity (Myerson, 1997).

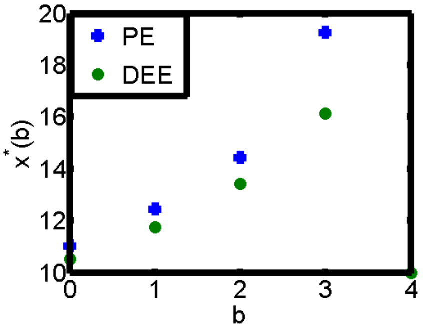

Figure 2: D's strategy at her preferred pure equilibrium (PE, shown with solid squares) and the dominant extend equilibrium (DEE, shown with open circles) for perfect monitoring with $s = 0.95$ and X following the uniform distribution on $[0, 20]$. Details are given in Example 6.

The following example illustrates the discussion on dominant extend equilibria.

Example 6. *Assume that* $s = 0.95$ *and* X *follows the uniform distribution on* $[0, 20]$*. We have seen in Example 1 that the maximum possible cutoff at a pure equilibrium is* $B^* = 4$*. At the best pure equilibrium for D, C trusts her when she has fewer than 4 black marks, and D's expected payoff is 60.18. However, there also exists a dominant extend equilibrium where C trusts D with probability 0.15 when she has 4 black marks, yielding an expected payoff of 65.07 for D.*

On the other hand, C's expected payoff at the dominant extend equilibrium is 0.15, which is significantly smaller than his expected payoff at any (non-trivial — with a non-zero cutoff) pure equilibrium. In particular, C's payoff ranges between 0.45 and 0.54 at the pure equilibria. Thus, if the truster is the one that chooses which equilibrium will be played, then a pure equilibrium will be chosen.

The values of $x^*(b)$ *are shown in Figure 2 for both the pure and the dominant extend equilibria. We observe that* $x^*(b)$ *is increasing and convex for* $b = 0, 1, ..., B^* - 1$*. We already know that this is true for pure equilibria by Proposition 1; a similar proof shows that this is always the case for dominant extend equilibria.*

References

Abreu, D., 1988. On the theory of infinitely repeated games with discounting. Econometrica 56, 383–396.

Abreu, D., Pearce, D., Stacchetti, E., 1990. Toward a theory of discounted repeated games with imperfect monitoring. Econometrica 58, 1041–1063.

Aperjis, C., Johari, R., 2010. Designing Reputation Mechanisms for Efficient Trade. Technical Report. Stanford University.

Athey, S., Bagwell, K., 2001. Optimal collusion with private information. The RAND Journal of Economics 32, 428–465.

Bhattacharjee, R., Goel, A., 2005. Avoiding ballot stuffing in eBay-like reputation systems, in: P2PECON, pp. 133–137.

Bolton, G., Greiner, B., Ockenfels, A., 2009. Engineering Trust - Reciprocity in the Production of Reputation Information. Working Paper Series in Economics 42. University of Cologne, Department of Economics.

Cabral, L., Hortacsu, A., 2010. Dynamics of seller reputation: Theory and evidence from eBay. J. of Industr. Econom. 58, 54–78.

Chwelos, P., Dhar, T., 2008. Differences in "truthiness" across online reputation mechanisms. Working Paper, Sauder School of Business.

Crawford, V.P., Sobel, J., 1982. Strategic information transmission. Econometrica 50, 1431–1451.

Cripps, M., Mailath, G., Samuelson, L., 2004. Imperfect monitoring and impermanent reputations. Econometrica 72, 407–432.

Dalea, D.J., Morganb, J., Rosenthal, R.W., 2002. Coordination through reputations: A laboratory experiment. Games and Economic Behavior 38, 52–88.

Dellarocas, C., 2005. Reputation mechanism design in online trading environments with pure moral hazard. Inform. Systems Res. 16, 209–230.

Dellarocas, C., Wood, C., 2008. The sound of silence in online feedback: Estimating trading risks in the presence of reporting bias. Management Sci. 54, 460–476.

Ely, J., Fudenberg, D., Levine, D.K., 2008. When is reputation bad? Games and Economic Behavior 63, 498–526.

Fan, M., Tan, Y., Whinston, A., 2005. Evaluation and design of online cooperative feedback mechanisms for reputation management. IEEE Trans. on Knowl. and Data Eng. 17, 244–254.

Farrell, J., 1987. Cheap talk, coordination, and entry. The RAND Journal of Economics 18, 34–39.

Farrell, J., Rabin, M., 1996. Cheap talk. The Journal of Economic Perspectives 10, 103–118.

Helland, E., Tabarrok, A., 2007. Does three strikes deter? A nonparametric estimation. Journal of Human Resources XLII, 309–330.

Lambson, V.E., 1987. Optimal penal codes in price-setting supergames with capacity constraints. The Review of Economic Studies 54, 385–397.

Mailath, G.J., Samuelson, L., 2006. Repeated Games and Reputations. Oxford University Press.

Miller, N., Resnick, P., Zeckhauser, R., 2005. Eliciting informative feedback: The peer-prediction method. Management Sci. 51, 1359–1373.

Myerson, R.B., 1997. Game Theory: Analysis of Conflict. Harvard University Press.

Myerson, R.B., 2009. Learning from Schelling's *Strategy of Conflict*. Journal of Economic Literature 47, 1109–1125.

Rosenthal, R.W., Landau, H.J., 1979. A game-theoretic analysis of bargaining with reputations. Journal of Mathematical Psychology 20, 233–255.

Schelling, T.C., 1980. The Strategy of Conflict. Harvard University Press, Cambridge, MA.

Shepherd, J.M., 2002. Fear of the first strike: The full deterrent effect of california's two- and three-strikes legislation. Journal of Legal Studies XXXI.

Sobel, J., 1985. A theory of credibility. The Review of Economic Studies 52, 557–573.

Tversky, A., Kahnenman, D., 1973. Availability: A heuristic for judging frequency and probability. Cognitive Psychology , 207–232.